AD 427161

DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA

SP-1492

How to Plot a Breakthrough

Lauren B. Doyle

12 December 1963

(SP Series)

SP-1492

How to Plot a Breakthrough

by

Lauren B. Doyle

December 12, 1963

SYSTEM DEVELOPMENT CORPORATION, SANTA MONICA, CALIFORNIA

ABSTRACT

Will there be a breakthrough in the field of information retrieval? One authority in that field has said, "No." This paper adopts the opposite viewpoint, and speculates on what the elements of such a breakthrough might be if it were to occur.

Several breakthroughs in other fields are scrutinized in order to highlight the factors which characterize and energize sudden expansions of new technologies. These factors, plus some factors specific to the field of information retrieval, are then extrapolated into a "plot for a breakthrough."

HOW TO PLOT A BREAKTHROUGH

"Breakthrough" is a word with military-journalistic origins which was apparently first used, in a large-scale way, to pertain to the breach of the German lines in Normandy during the summer of 1944, and the subsequent race by Patton and other elements of the Allied armies to Paris and points east. Since the war, the word has been repeatedly used to refer to a phenomenon which has become especially frequent in the last two decades--the "technological breakthrough." In analogy to the military breakthrough, there is the entrenched enemy of a formidable technical problem, the breach made by some powerful new approach or technique, and the exploitation spearheads, which may fan out in many unexpected directions.

As is usually the case with glamour-infused terms, vulgarization sets in, so that anybody and everybody can benefit from the "psychological fall-out" of the usage of the term. When so many people use a word like "breakthrough" in connection with promotional efforts, the use of the word soon becomes suspect, and those few who still employ the word in an honest way are often themselves forced to find still another word to describe what they are talking about.

Nevertheless, the word "breakthrough" is just as much here to stay as is the phenomenon to which it refers. Furthermore, if we understand what a breakthrough is, and why and how it takes place, we are in a much better position to use the word meaningfully and to shape research and development efforts in ways to increase the speed of incipient breakthroughs as they are recognized. One can go even a step further and assert that, knowing all the major characteristics and possibilities of a given problem area, one can actually <u>plot</u> a breakthrough. It is admittedly a tricky thing to do--and failure is overwhelmingly probable--but success <u>is</u> possible.

What are the defining attributes of a breakthrough? As we enumerate them here, in each case we are going to look at some pattern of events in history which displays the enumerated attribute along with other typical breakthrough attributes.

The first attribute is the <u>relative shortness of time</u>[1] over which a breakthrough materializes and grows to maturity. This can be illustrated by what was perhaps the first genuine breakthrough in the history of man--the control of fire. Man is said to be a tool-making animal, but his development in the use of tools proceeded with painful slowness over tens of thousands of years; the use of tools by humans could be called a breakthrough only in the perspective of geological time.

[1] This attribute is descriptive, whereas those to follow are contributory or causative.

for electricity was latent--but perceived and understood by a large enough number that the breakthrough was practically obliged to happen.

The latent need constitutes only half of the driving force behind a breakthrough; the other half comes from the emergence of clearly superior new methods, the third attribute of a breakthrough. Arguable superiority is not enough; superiority must be incontrovertible. In the contest to which we have just alluded, electric light versus gas light, the problems of constructing untried and unfamiliar electric power distribution networks were so great that such enterprises might not have been attempted except for the fact that, once the electric light had been shown feasible, almost anyone could see that the use of electricity was superior in almost every important respect to the use of gas, even when a battery was the electric power source.

In modern times one of the best illustrations of this principle is found in the transistor. In the ever-shorter time spans in which breakthroughs are taking place, we find less than a decade has elapsed between the first public announcement of the transistor's invention and the large-scale manufacture of transistorized computers and other devices; and today, after only fifteen years, transistors are the basis of a billion-dollar industry.

In part this rapid progress was energized by the requirement for miniaturization and low power in airborne and satellite computers. But the transistor, quite apart from the special needs of modern weapons systems, has such general superiority over vacuum tubes that a breakthrough could have been expected in any event, once the characteristics of the transistor were made known to the technological community.

The vacuum tube is fragile, cumbersome, and wasteful of power; its fragileness leads to unreliability and to sensitivity to shock and vibration; its cumbersomeness requires complex instruments such as computers to be bulky and spread out over entire floors, like the stacks in a public library; its high consumption of power leads not only to unnecessary operating expense, but also creates added engineering problems connected with heat removal, especially when much electronic equipment is required to be in a confined space, as in an aircraft or aboard ship. But the transistor has none of these disadvantages.

The transistor, then, is one of the more remarkable examples of across-the-board superiority of a new technique over an old one that makes a breakthrough "in the cards." How many other examples are found, in history, of rapid technological expansion following widespread realization of superiority of method? It would be difficult to cite them all. There are irrigation (3000 B.C.), movable type (1450 A.D.), steam power (1770), railroads (1804), telephony (1876), radioactive tracers (1934), amplification by stimulated emission (masers, lasers, 1951). These are only a few of the more spectacular examples.

A fourth attribute of a breakthrough is adequacy of available supporting technology. Perhaps there was a need for automatic computational machinery in 1823; perhaps Babbage's method of mechanical computation was clearly superior to anything in sight; unfortunately, the supporting technology required for the feasibility of Babbage's system did not yet exist.

There are other less dramatic examples of situations where the need and the superior method were both available, but where the actual breakthrough was postponed until the required supporting technology was developed. Though Lee de Forest invented the three-element vacuum tube which made radio reception feasible in 1906, radio broadcasting did not come to pass until 1921; manufacturing methods were just not good enough prior to that time to assure reasonable reliability of performance in an electronic device even as simple as a radio.

The importance of "adequacy of available supporting technology" is not to be underestimated merely because it is given here as the fourth attribute, rather than as one of the first three. There have been times when either the reality of the need or the superiority of method was not yet clear enough to be seen by those who needed most to see it, but where supporting technology together with a handful of imaginative entrepreneurs turned the tide toward breakthrough.

When intercontinental rocketry began in the early 1950's, the accent both in the U. S. and Russia was on liquid-fueled systems; solid-fueled rockets were thought of as nothing more than glorified artillery shells. The U. S., however, had the good fortune to have a highly developed organic polymer technology, which included synthetic rubber and plastics. Some of the organic polymer industries saw that there was a future for them in solid propellants and began, more or less on their own, to plot what would eventually become a breakthrough.

The major problem was to influence the people who held the pursestrings in the Defense Department to finance the more expensive phases of development of large solid-fueled boosters, and it was at this point that the "clear superiority of method" slowly began to assert itself. Even here progress was difficult, because those factions with vested interests in liquid-propellant system development were most persuasive in their arguments to the effect that solid fuels were inherently inferior. One problem was the seemingly inadequate thrust of solid fuels. Another was the problem of cutoff control; stopping a solid-fueled rocket when the proper velocity was achieved seemed about as difficult as stopping an erupting volcano.

As we now know, problems of the above sort were not insoluble, but required only a reasonably steadfast application of engineering know-how and ingenuity. In Russia, with its somewhat underdeveloped petrochemical industry and its trailing position in the development of compact nuclear warheads, the case for solid-fueled intercontinental missiles was never strong enough--during the 1950's--to justify the developmental investment.

Doubtlessly most of us remember the first newsreel shorts of a Polaris missile being fired from its submerged mobile launching pad. What we saw was slightly incredible: a ridiculous "pop bottle" thrusting itself out of the ocean, drunkenly igniting itself, and taking off like a Fourth of July rocket--in great contrast to the agonizing slowness of the initial flight of an Atlas or a Titan. No one could then deny that a breakthrough had achieved maturity.

As preparation for discussing "How to Plot a Breakthrough," we review the attributes so far discussed, with an indication that each must be taken into account if the plot is to succeed.

1. Relative Shortness of Time. This attribute was presented first because it describes what our goal is: to bring to a state of widespread application a new technology whose principles are understood at the outset by perhaps no more than half a dozen people, in a period of time which is quite short in comparison with typical technological gestation periods. Commercial atomic power is an example of a non-breakthrough in the sense that, though the principles of technology were understood in 1942, essentially twenty years were required to achieve the state of "widespread application." In today's times a self-respecting breakthrough should take no longer than five years.

2. Latent Need. This factor was presented second because, though it is a driving force for a breakthrough, it is a most difficult force to harness. If one attempts to "plot a breakthrough" and fails, it is very likely because he does not understand how to assess and capitalize on the need factor. A technical person may have the educational qualifications and experience to see where there is a breakthrough to be made--and even be right--but be fatally unaware of the quasi-political processes by which strategically placed people become convinced that the need does exist and can be met through a new technology. Unless more than a few are induced to develop and apply the new technology, the support required in the early phases of breakthrough cannot be mustered. This decade sees so many "competing breakthroughs" and so many high-pressure types telling consumers--big and small--what they need, that a breakthrough without an adequate public relations apparatus is headed for stagnation before it starts.

3. Clearly Superior Methods. We think of a man as fortunate indeed if he stumbles onto a method or technique which looms head-and-shoulders in effectiveness over prevailing arts. But it is not necessarily a case of luck; such techniques can be deliberately sought and found (this is why we finance what is known as "research"). The reason so few are rewarded in the search is not so much that most people are unlucky, but that most people do not persist long enough in the search--they are content to invest their time and energy in the first marginally superior idea they encounter. To put it another way, if one aims for Mars and in the process reaches the moon, he becomes satisfied with the lesser goal because he never dreamed that he would be one of the privileged few to stand on the moon. The avenue, then, to finding clearly superior methods is: don't be satisfied with the moon.

4. Adequacy of Available Supporting Technology. For the technically oriented person this is perhaps the easiest factor to reckon with in the plotting of a breakthrough. But, of course, its importance nevertheless is not to be forgotten for a moment. Indeed one might begin his breakthrough plot by searching for brand-new supporting technologies which have not been around long enough for people to have realized their most important ramifications. Most of the people who are aware of the power of such a technology are often themselves leanly supplied with ideas for its exploitation; typical behavior is to apply the technology on a hit-or-miss basis hoping something good will happen. As is frequently the case, such technologies are often so potent that even a random, unthinking attempt at application can sometimes result in success. If this is so, the prognosis for a carefully planned applications foray is good indeed, assuming of course that the plan is flexible enough to take advantage of unforeseen developments.

In which field shall we now plot a breakthrough? Don R. Swanson ((1)) has said in his 1961 Western Joint Computer Conference talk, "Information Retrieval: State of the Art":

> "...Semantics and redundancy are key conceptual issues and give rise to difficulties more likely to be overcome by meticulous thesaurus compilation than by any sudden insight or "breakthrough."..."

To a breakthrough plotter such a statement can only be regarded as a challenge. It reminds him a little of the 19th century statements that since everything is already known about the physics of matter and energy, future progress is merely a matter applying the known to practical problems. The phrase "meticulous thesaurus compilation" is particularly odious because it seems to suggest that digital computers, far from rendering the retrieval process automatic, actually make it necessary for people to do more work.

Our first task, in plotting this particular breakthrough, is to inventory available supporting technologies, old and new. First, the old. We have 1) library methods and 2) the publishing industry. It is not immediately apparent that traditional libraries have much to offer us, aside from a certain history of intellectual grapplings with problems of classification and subject indexing. The publishing industry, however, contains many resources which could play a part in a breakthrough. We center our attention especially on the increasing tendency of publishers, for various reasons, to handle textual material in digitalized form. There are teletypes, monotypes, flexowriters, and other digitalizing instruments which convert vast volumes of text into a medium wherein it is available without further human processing for storage in computer memories.

The second thing of importance about the publishing industry is that it has a large cadre of people who are accustomed to working in a highly mechanized environment, where many of them are typically middlemen between one

machine and another (examples: from teletype to copy editor to Linotype, or from Linotype to proofreader to press).

We now look at new technologies which are available and (potentially) adequate. Chief among them is the digital computer, with all its varied peripheral equipment. We wonder: given the power of the digital computer, why is it that a breakthrough has not occurred already? The hardware is adequate. The software (programming systems and languages) is also adequate. Wherein lies the barrier? It is probably in the lack of anything in the province of "clearly superior methods"—what might be called the "ultra-software" needed specifically in dealing with situations involving the retrieval of natural language text. The ultra-software technology is not yet adequate, and is not among the supporting technologies we can count on.

A second available new technology centers around cathode-ray tube displays. There are a variety of existing ways by which text can be displayed on scopes and, say, corrected or revised by an editor without ever having to go through the output-correction-input cycle which is so cumbersome with the usual in-out techniques. These ways range from instruments such as Aeronutronic's Flidex, in which outgoing teletype messages can be displayed to the operator for correction before actually being sent out, to more advanced scopes accepting light-pencil input.

A third new technology tends to supplement that already existing in the publishing industry; in addition to the various now-used machines which digitalize text, we have the photoelectric print reader. Its potential is great because it renders text input as easy for the computer as the human eye renders it easy for the brain. Unfortunately, print readers are not yet able to accept any and all printed material placed before them, because of the bewildering variety of type fonts which exist in modern printing, and therefore because of the difficulty of developing character-interpretation logic as generalized as that of the human eye and brain.

If we are now to have any real chance of finding a "clearly superior method," we must survey the latent need as well as the adequate available supporting technologies. It is a great deal harder to assess needs in a field like information retrieval than it is to inventory available technologies; the open professional literature describes the latter in as much detail as we are ever likely to ask, but the former are described in a way which could hardly be anything but superficial. This is not because we are poor observers of human needs, but more because there are so many important elements, even critical elements, which are either extremely difficult to observe or actually unobservable.

You can't rely on what people say they need; you can't even draw definitive conclusions from what they do use. How could we have predicted, for example, the effects of telephones and automobiles on consumers, in advance of their

invention. Alexander Graham Bell was only the first to discover the telephone—every telephone user since then has "discovered the telephone" in his own private world, by finding out how to integrate the telephone into his own pattern of living. But if one had asked a resident of a New England town in the 1870's whether he needed a telephone, the answer would probably be something like: "Look, Mister. There's too many people talkin' as it is. A body can hardly get a day's work done without some good-for-nothin comes around to pass the time jabberin' about his sick cow. The last thing I want is that infernal bell ringin' right in the middle o' horse feedin'."

The approach to need-assessment with the greatest probability of success is the one which observes how people do actually discover the utility of both new and old facilities. Self-observation can be of great importance, because the opportunity of noticing significant things is ever-present; this is especially true in an area like information retrieval, where people interact with the written record as individuals, and where the observer has many more chances to monitor his own information-searching forays than to monitor those of others.

One of the first things a conscientious self-observer will become aware of is that there is a relationship between the frequency of his searches and the distance between his normal work location and the medium to be searched; in effect there is an inverse proportion between frequency and distance.

The shortest distance to an information source occurs, of course, within his own head. If the information required is not easily accessible in the mind, one then looks for an information source within arm's length of his chair: his bookshelf, his notebooks, his correspondence file, his stack of professional journals, or—if need be—the telephone. The pile of paper in one's office is in effect an auxiliary memory with a typical access time of one minute. A person may work the whole day and not have to make more than a dozen forays beyond the confines of his office.

Out of this dozen, fully nine or ten might be to neighboring offices within a radius of 50 feet or so of his desk, either to talk to colleagues or to consult books or journals known to be in their possession. Perhaps, then, only on two or three occasions during the working day is one required to venture beyond his immediate neighborhood in search of information. As a conscientious self-observer I notice myself visiting the company library not more than twice a week. Other "distant" forays are typically: 1) to consult books which I keep at home, 2) to write letters to individuals known to have specific information, 3) to peruse special (non-library) collections of documents possessed by more distantly located colleagues, in and out of the company, or 4) to attend professional meetings.

In addition to revealing the inverse proportionality between frequency of search and distance to search media, self-observation also reveals that one

gets much advantage from knowing where information is to be found. I have made several abortive attempts to construct and maintain a card file index of the contents of my office, but found literally that it was not worth the effort expended. Unwittingly, and effortlessly, I was maintaining in my head a better index of the contents of my office than I could construct by hand. Aside from easy preparation and easy access, the mental index has the advantage of not requiring fussy decisions over indexing terminology, or under what heading or how many headings should a given item be filed.

The "need analysis" we have made so far points to an interesting new suggestion for improving the average man's access to information: to improve the quantity and quality of the information-search aids within arm's length of the information consumer's position at his desk. This policy accomplishes two things: 1) It permits a large number of bits of "information about information" to come into the consumer's swiftest access auxiliary storage; his lack of inhibition in consulting these sources would be second only to his lack of inhibition in interrogating his own mind. 2) Because of the high frequency with which the consumer will consult the sources within arm's length, he will form a "mental index" of those sources which in time will greatly increase his efficiency in their use.

Our hypothesis, then, is that the information consumer's greatest latent need is for one-minute-or-less access to highly organized, information-rich material which gives him maximal contact with that portion of the world's written record pertinent to his work. Unless someone can suggest an alternative need-hypothesis, we will adopt this as a major premise in the plot to break through.

The next step is the quest for a clearly superior method. We know the available technologies. We have a reasonable hypothesis of need. How can they now be brought together to increase the mental contact between the information user and the written record?

We must now remind ourselves anew of the most likely reason why a plot for a breakthrough will fail--the difficulty of harnessing the latent need as a driving force for the breakthrough. Let us suppose, by way of illustration, that one conceives a computer-produced index which is clearly superior to any other index--manual or automatic--in sight. How would such an innovator proceed to convince the world of users of the clear superiority of the index? The most direct way would be to circulate the index to a large number of users; because if the superiority of the index is clear, it ought to be especially clear to one who uses it.

Unfortunately, much developmental work and capital investment is required before one can ever circulate such a product to a large enough number of people to create a "self-sustaining breakthrough." The inexpensive alternative is to convince strategically placed persons, who control dollars and man-hours, that the index is clearly superior. At this level, however, superiority may

not be so clear as it would be for a user of the finished product. Because these people have greater responsibility, they also have greater caution. Moreover, they would be constrained to evaluate the index according to intellectual criteria rather than according to personal experience in usage. Furthermore, an undeveloped idea is never as attractive as a finished product, and a man is especially suspect when he claims his idea is "clearly superior."

This is a fact which has to be contended with: ideas which later prove to be decisive too often are unrecognized and/or opposed in their infancy. But one thing must be recalled from our very first sentence about latent need: that such need is accompanied by a receptive climate where unusual welcome is extended to new ideas--the difficulty is that, in modern times especially, the hospitality of this climate is abused by a saturation with many ideas which are not so good; in other words, the climate can be so receptive that bad ideas are frequently purchased, causing receptivity to become weighted down with skepticism and caution.

Receptivity, however, is still extant, and if we can locate its surviving foci we can mobilize the psychological capital, and eventually the budgetary capital necessary to give the breakthrough momentum. The most receptive possible ally is one who already has a technology packaged up and for sale. Just as the polymer industry was looking for solid-fuel rocketry, so might some present-day industry be looking for the clearly superior retrieval idea. If they can be satisfied as to the clear superiority of it, they will join the breakthrough plotters energetically in developing it and promoting it.

Now, assuming for the moment that there does exist some benefactor with a "technology for sale" who will support any "clearly superior idea" which we might see fit to advocate, and for the moment not worrying about who this benefactor might be, we now feel free to uncover the clearly superior path to retrieval of documented information.

What should we put within "arm's reach" of the information user at his desk? How should we fill his "quick access auxiliary storage"? The answer which suggests itself is: we should surround the user with highly organized and condensed summarizations of the literature. They must be handy enough, palatable enough, and rewarding enough that the user will consult them frequently, because the more frequently he consults them the better his "mental index" will become, thus in turn lowering his inhibition in consulting his "arm's length store" and increasing his effectiveness as a retriever. Handy enough, palatable enough, and rewarding enough--we take up each of these requirements in turn, in the hope of establishing the specifications for the clearly superior retrieval instrument.

Handy Enough. Part of the handiness is, as we've realized, arm's length proximity. Books, pamphlets, or file cards are handy. Microfilms, and other such ultra-condensed representations, allow us to place more bits of information within arm's reach of the user, but at a double cost: 1) some increase

in use inhibition, and 2) much less feasibility of "mental indexing," requiring recourse to an external indexing scheme with all its time-consuming fetters. Microfilm, of course, could serve as a somewhat larger and less accessible office storage, and if a correspondence were to exist between book or file-card (macro) storage elements and the more capacious elements of the microfilm, then the "mental index" giving access to the macro-storage will also serve in part as an index to the microfilm store. This is optional, however, and it appears a reasonable stance that the "arm's length store" should be composed of macro-elements.

Palatable Enough. The elements of the user's arm's length store should be easy to interpret, and preferably even fun to interpret. A large part of our success in having any retrieval system work is in persuading people to use it frequently and voluntarily. If there is joy in using a system, then little persuasion is necessary. It is a sad fact both for traditional reference systems (bibliographies, abstract journals, indexes) and for computer-generated products (auto-abstracts, permuted title indexes) that their use is a dismal way to spend one's time. In a pre-computer society this might have been inherently unavoidable. We must now ask: can computers help us to generate reference systems which are a pleasure to use?

Rewarding Enough. Many otherwise uncriticized retrieval schemes have often fallen flat because a user had great difficulty in obtaining the documents to which the system referred him. Generally speaking, in any system we must expect the document to be less accessible than its condensed representation. However small the access time and the inhibition in the user's reference to his arm's length store, the purpose of the whole idea is defeated if each search trail leads simply to a reference to a document. This is not so merely because the document might take an hour or a day to deliver, but even more importantly because the user may not want to read a whole document in each case. In short, our arm's length retrieval system must provide information rewards of a brief and crisp character, so that the user is given <u>some</u> information whether or not he orders the document. Thus, even if the user <u>never orders a document</u>, he will still derive enough value from his arm's length system to motivate him to use it frequently.

The formulation which now suggests itself is to have computers generate, from the full texts of documents, condensed representations which can be combined in an organized way and issued as books, pamphlets, or card sets available on request (or through selective dissemination) to users for use at their desks.

There are many aspects of such a system which we do not have space here to discuss. For example, there is purging and updating to be considered; we have room only to opine that the user should be given the option as to whether these functions are under his personal control (requiring added effort on his part) or under some form of automatic or institutional control.

The only aspects of the "arm's length system" which we want to discuss at length in this article are those germane to our attempt to consciously plot a breakthrough. We have just decided what the focus of the breakthrough is to be: a highly organized, easily accessible, information-rich auxiliary storage within arm's reach of the information user at his desk. The next questions are "What will the elements (books, pamphlets, card sets, etc.) of this store contain, and how will they be produced?"

What we first must think about is organization. The arm's length system has two basic levels of organization, gross organization (arrangement of the elements) and detailed organization (arrangement of sub-elements within each element). On the gross organization level, we have appreciated that the desired low inhibition in use of the system can best be realized if the user is allowed to arrange them willy-nilly and to rely on his own familiarity with where things are in his office environment. If the user is a compulsive type who likes to carefully and deliberately organize and index his environment, this should be his prerogative. Our system, if it is to be truly popular, must accommodate a variety of temperaments.

We assume that most users, compulsive or otherwise, will unavoidably rely heavily on their "mental index" of their auxiliary store; this leads to the requirement that there must not be too many elements in the store; it argues that books are preferable to pamphlets, which in turn are preferable to cards. On the other hand, books are bulky, relatively expensive, and not amenable to updating and selective purging. A tentative "optimum" might be packets of pamphlets, where pamphlets can be grouped according to the user's own conception of how the universe should be organized. The physical means of grouping could be box files, binders, vertical partitions, or whatever means might appeal to a user.

As we go from the gross organization level to the level of detailed organization, we also travel from a realm where it is easy for the user to arrange things to suit himself to a realm where there are far too many units to be arranged by the user. As we go from the large to the small, the "mental index" principle becomes less and less effective, and the willingness of the user to organize things declines rapidly. And yet at the detailed level organization is of critical importance.

How are we to achieve organization at this level? At this point we have recourse to one of the most mighty of our available new technologies: the software technology known as "digital computer programming." By a variety of ways we can feed the text of entire articles into computer storage and, by programmed means, perform a practically infinite variety of operations on the stored text.

One of the operations we can perform is to group documents according to similarity of word content, which approximates to a reasonable degree grouping

by topical similarity. There are working programs now available which can do this, although the art of "automatic grouping" is just beginning and cannot be called an "available technology" in itself. Nevertheless, it is becoming apparent that pushing this art is an important component of the breakthrough we are plotting. It is not, however, an indispensible component; there are grounds for believing that if the software available for "automatic grouping" remains in its present primitive state it could still serve as a factor in the breakthrough.

We need not be unduly disturbed if similarity of word content is not in precise correspondence with topical similarity. For one thing, "topical similarity" is hard to judge when we deal with articles in the same field. There are too many arbitrary subjective criteria by which one might decide that article A is more similar to article C than it is to article B or vice versa. For this reason, matching of choices of words by the authors of the articles is as good a basis as any for determining degree of similarity. Furthermore, this method of determining similarity has the indisputable advantage of lending itself to automaticity.

Now, numerous researchers in the field of information retrieval have studied document grouping according to similarity of word content, but no breakthroughs have occurred in consequence. Why not? Conscientious as the studies might have been, the importance of representing the groupings in a way that is visible to and interpretable by the information searcher has not been seen. All systems advocated on the basis of these studies have had high inhibition factors. When we see that our approach to retrieval must involve low inhibition coupled with enhanced motivation, we are on the road to a "clearly superior method."

Our first tentative decision was to provide the information user with packets of pamphlets for the user to arrange as he sees fit. Our second tentative decision is that the organization of the material in the pamphlets should in some way be based on word-content similarity as determined by appropriate computer programs. What this means in particular is that all of the documents represented on, say, page 108 of a given pamphlet should bear a closer resemblance in word content to each other than they do to documents represented on other pages. (Such a state, by the way, does not have to be literally achievable; similarity relations do not necessarily structure themselves in such an obliging way. All that matters is that it be "approximately achievable.")

We are now ready for a more detailed description of a typical pamphlet: what all is in it and what does it look like? Like any reasonable pamphlet, this one will have a "table of contents" at the beginning; the table will be in large part computer-generated; but it will also be <u>heavily edited</u>. Without the latter operation, the breakthrough cannot get off the ground, in this decade at least. We must not forget that "palatability" is an

important factor in enhancing the motivation of the user to frequently consult his auxiliary storage. Unedited computer output is likely to be unpalatable for a long time to come.

After the table of contents we come to the basic material of the pamphlet. Each page will have a dozen or so representations of topically close documents. What will they be like? Will it look like an abstract journal? If it is literally like an abstract journal, we can bet that the breakthrough will not go; I know professional people who have in their office cartons full of abstract journals which they never look at.

Part of the trouble is the "underorganization" of these journals; but also it is, I think, that there is something repulsive about a page full of abstracts. We must construct a page which will draw the user in. In the center of the page (or at the top) will be a large word or term, in boldface type $\frac{1}{4}$-inch high. It will be a word or term suggestive of what the documents represented on that page have in common; it will very likely be a content word (or term) which all of the documents have in common as a key word, i.e., a word of high frequency. Radiating outward from the centrally placed word, thin lines will lead to words in smaller type which reflect what certain sub-groups of the documents have in common. Finally, for each document will be shown a single sentence--a conclusion, fact, or opinion.

Would this sentence be picked out of the document automatically? Largely yes. The actual process would probably be something like this:

1) A computer program selects six or eight sentences from document text, based primarily on richness of word content. Ways of doing this have already been worked out (2,3).

2) An editor selects one of these and rewrites it so that it can stand by itself, without requiring the context of the document for its easy interpretation. In an advanced index-generating system this could be accomplished at a display scope with a light pencil.

3) The sentence, once designated by the editor, is assigned by the computer program to that page of the index which contains its closest topical neighbors.

There are, to be sure, dubious aspects of this scheme. The authors of the documents may be upset at the thought of having a single sentence of theirs, out of context, displayed to thousands of information users of whom only 1% might ever order the entire document. I am sure, however, that these same authors would feel no pain upon reading similar sentences drawn from the text of other authors. If there is uncertainty about the truth or significance of the sentence, one can always order the article. The fact is, in

most cases it <u>will not matter</u>. The world is already constructed in such a way that the user's picture of 99% of the documents to which he is exposed is fragmentary indeed. Furthermore, as a system like that just described undergoes evolution, the authors may themselves be allowed to designate or specially compose the sentences which will represent their articles. Another dubious aspect of this scheme is that it sounds like a tremendous amount of human labor would have to be put into the index-generating process if an editing operation is required for every entry in the pamphlet. On the other hand, the editor pictured in our scheme has to do far less work per document than a conventional abstract writer would have to do. If, through computer use, we make a 90% saving in editorial effort, it is almost as large as the 100% saving we would make under complete automation of index generation. The question we have to ask is: what does the residual 10% of editing effort accomplish?

In this case it accomplishes a great deal, because without it we would not have a breakthrough. The requirement for "palatability," as we have already noted, is crucial enough to justify a considerable investment in editing. The investment in editing is acutally but a small added cost which we regard as an insurance premium--that the millions of computer program operations required to analyze, compare, group, organize, and condense documents will not have been expended in vain.

What else is in the pamphlet? In addition to the "palatable" portion of the pamphlet, to which the user will refer most often, there will be a "business" portion, which need not be palatable and can be automatically generated without a requirement for editing. The "business" portion will contain two sections. One will be a standard bibliography containing all of the information the user would need in order to find the document itself. The other section would list pamphlets containing more detail per article, in case the user elects to have a more detailed index of an area of special interest to him. Finally, there is an index, generated automatically with perhaps some editing, which will provide alphabetical entry to the pamphlet, thereby supplementing entry on the basis of topical similarity.

Now we take a second look at the "arm's length system," in quantitative and qualitative terms. Quantitatively, we see perhaps 200 pamphlets within easy reach of a user, each of which contains references to perhaps a thousand documents: 200,000 "palatable" references in a highly organized state. Such a system could well accommodate references to all of the articles in one's own profession for the past decade, plus a good number of references in related fields of the user's choice.

Qualitatively, what do we have? What it amounts to is a highly structured collection of proverbs. Each such proverb (which need not necessarily be factual) is the crisp "information reward" which we decided earlier would be required to enhance user motivation. There is then not only low inhibition

in using the file, but also a rather strong temptation to browse. Many amusing surprises would lurk in the file, such as "The Zwirtz Process requires phospate rock feed with low silica contamination" next to "The advantage of the Zwirtz Process is its tendency to precipitate and discard silica impurities." Or we might see, "It is unthinkable that we can have automatic indexing without first developing machine searching" adjacent to "From a research viewpoint, a study of automatic indexing is an essential prelude to the solution of the machine searching problem."

To reiterate: when searching the literature becomes convenient and entertaining, people will search the literature. Indeed, five years from now it could be commonplace to hear supervisors saying, "Jones, will you please stop searching the literature and do some work."

The question we finally ask is: Who will provide the money and the man-hours to start this breakthrough on its way? Who is the benefactor with the "technology for sale" who will support a "clearly superior idea"? It is highly likely that such a party will be found within the publishing industry. The motivation of this party would be pure and simple: Is this "clearly superior idea" a good way for me to make money?

The possibilities for market expansion are large. Consider the following quotation of John Markus (4) of the McGraw-Hill Book Company:

> "...Money is at the root of most index-publishing problems. Individuals rarely buy published indexes, partly because of their high and continuing cost and partly because they usually have to go to a library anyway to get the documents cited in an index. The market for index volumes is thus small because it consists largely of libraries...."

When we talk about hundreds of pamphlets lining the offices of tens of thousands of professional workers, we are talking about a huge market which does not now exist. The major question of interest to the publisher is: What will it cost to produce these indexes and will people buy them? Questions such as "Is similarity of word content a sound basis on which to organize references?" would tend to be of little interest to the publisher; it would be of much more interest to him that automatic organization _of any kind_ is feasible, and here again the question would be "How much will it cost?"

Many publishers will find themselves already having, as part of their normal operation, most of the text they publish in digitalized form. This puts them in an excellent position to develop a pilot-plant operation among their own clientele. Expenses for computer time, which in the scheme we describe are still quite large, might be willingly borne--as part of a development effort--if cost projections for future computer expense rates decline steeply enough. And this is not the only possible source of cost

lowering. Another very substantial decline could come from finding programmed methods vastly more efficient than any we now visualize; but one couldn"t hope to realize such gains without being willing to undertake a development effort.

A publisher, furthermore, would tend to be uninterested in such questions as: "Is machine searching of text superior to machine indexing?" If he can sell pamphlets to someone who works twenty feet from the output end of a literature-searching computer, that's good enough for him. He would not be interested in the outcome of rigorous evaluation experiments which determine whether Index A is better or worse than Index B. But if Index B sells faster than Index A, he is then very much convinced that Index B is better. The publisher is not alarmed when he reads that the number of literature producing-and-consuming scientists is increasing exponentially. Indeed, he is quite pleased, because it implies for him an exponential increase in business volume.

The publisher would not take part in earnest debates about whether or not we should have mammoth centralized information centers, as the Russians have. He would instead take comfort in the realization that his customers would be on the average several hundred miles from any such conceivable information center. When an information center is finally opened, it could only mean to the publisher a more comprehensive supply of raw material from which to generate pamphlets.

If publishers do indeed decide at this time that there is an "information retrieval market," from which money is to be made, there are some interesting evolutionary possibilities. One such possibility, for example, is that the balance between man and machine in the production of indexes may shift back toward the man, rather than toward the machine. In a doctrinaire climate where people are satisfied with nothing less than "fully automatic high-quality" processing of language, such a trend could not occur. But it could occur if a publisher found that increased application of editorial skill resulted in increased sales.

The question of "How to Plot a Breakthrough" cannot be answered in the text of this paper. It is much easier to write about plotting a breakthrough than it is to plot one and have it thereupon take place. In one sense it is about as foolish as writing about "How to Make a Million Dollars" without having first made it. Indeed, even one who actually makes a million dollars is not necessarily able to give the public the straight goods on how to make a million. If such a millionaire were in full possession of the truth about human nature, and were honest in revealing his knowledge, he might say: "Don't even try to make a million. There are only a few people who can do it, starting from scratch, and I'm one of them."

Therefore, even if I had already successfully plotted a breakthrough I would not be by that token in a position to advise others on the matter. On the

other hand, the theme has been emphasized herein that if breakthrough plotting is at all possible, it can only be accomplished by studying former breakthroughs and understanding the forces behind them. If one then looks for equivalent forces in a current problem area, he stands a good chance of finding them.

Spotting these forces, of course, does not automatically lead to a successful forecast. One can leave an unrecognized important element out of his calculations--an element which can lead history in a completely different direction than one might predict. Maybe there is some unrecognized virtue--for example--in not keeping up with literature. I have found that many professional people are intimidated in their productivity by their over-awareness of how much the rest of the world knows. Will more effective contact with the literature only intimidate them more? Do they intuitively know this, and do they therefore try to avoid the literature? This paper has tried to include the probable psychological characteristics of the information user in its analysis, but it may not have gone nearly far enough.

Whatever the case, it is a safe statement that practitioners of research and development in the field of information retrieval have steadfastly neglected to appreciate that the mind and psychology of the information user form the central element in the whole retrieval picture. Some grudging acknowledgments of this have been made, leading to library use studies and retrieval system evaluation projects. These studies, of course, shed very little light on what people will do when given radically new tools, such as modern technology is making available. If I may be so bold as to say it, perhaps our studies of the user are not sufficiently "fundamental." One is hard put these days to win support for such research, because it is "too abstract" and "uninteresting." It is "too remote from practical application." And for some strange reason the horizon of "practical application" continues to be bleak. The "fundamentals" of human nature, however, will catch up with us whether we know them or not.

This could be a large part of the story of why breakthroughs happen. Do they happen suddenly because they are long overdue? The principle of lasers was described by F. G. Houtermans, a physicist, 30 years ago. Why was there not a gradual development, rather than a sudden breakthrough? Solid-fueled rockets are as old as gunpowder. Why didn't the Germans (who once led the world in the technology of organic chemistry) develop Polaris-like missiles? Electricity and its properties were known about in the times of Napoleon. Why did it take a century to reach the "age of electricity"?

The balloon of dogma and mental rigidity is inexorably expanded by the pressure of events. The tension increases. Then along comes one guy with a pin. The result: poof! Another breakthrough has occurred.

REFERENCES

1. Swanson, D. R. "Information Retrieval: State of the Art." Proceedings of the Western Joint Computer Conference, May 1961, pp. 239-246.

2. Luhn, H. P. "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, April 1958.

3. Doyle, L. B. "Indexing and Abstracting by Association," American Documentation, 13 (4), October 1962, pp. 378-390.

4. Markus, John. "State of the Art of Published Indexes," American Documentation, 13 (1), January 1962, pp. 15-23.

UNCLASSIFIED

System Development Corporation,
Santa Monica, California
HOW TO PLOT A BREAKTHROUGH.
Scientific rept., SP-1492, by
L. B. Doyle. 12 December 1963, 21p.,
4 refs.

Unclassified report

DESCRIPTORS: Information Retrieval.

Discusses a breakthrough in the field of information retrieval. Speculates on what the elements of such a breakthrough might be if it were to occur. Scrutinizes several

UNCLASSIFIED

UNCLASSIFIED

breakthroughs in other fields in order to highlight the factors which characterize and energize sudden expansions of new technologies. Reports that these factors, plus some factors specific to the field of information retrieval, are then extrapolated into a "plot for a breakthrough."

UNCLASSIFIED